Institute of Technology, Nirma University
DM Innovative Assignment
2CSDE71

Group:

2OBCE307 - Mufid Vohra

20BCE323 - Yati Tank

TOPIC:

Crop price predictions using the given dataset, which reflects price area, and other attributes.

DATASET DESCRIPTION:

The dataset provides total information of 23473 surveys dating as back as 1866. All surveys were done state-wise, and we will focus on the corn grain. It is summarized by a value for each study corresponding to the state.

The second dataset contains features for various crops and we have applied the best algorithm to make sure most accurate predictions for the same.

NOVELTY:

We have used one hot encoding here. One hot encoding is a technique used in data preprocessing to represent categorical variables as numerical values, which can be used as input for machine learning algorithms.

We have implemented this on 3 data frames from the dataset. Period, Data Item, and State.

We have also compared the metrics for all models possible, to find the best fit according to our dataset, and can conclude that Random Forest Regressor gives us the best R2 value.

Following up on this we used the same random forest regressor on a larger dataset containing different crops, to make sure it follows through.

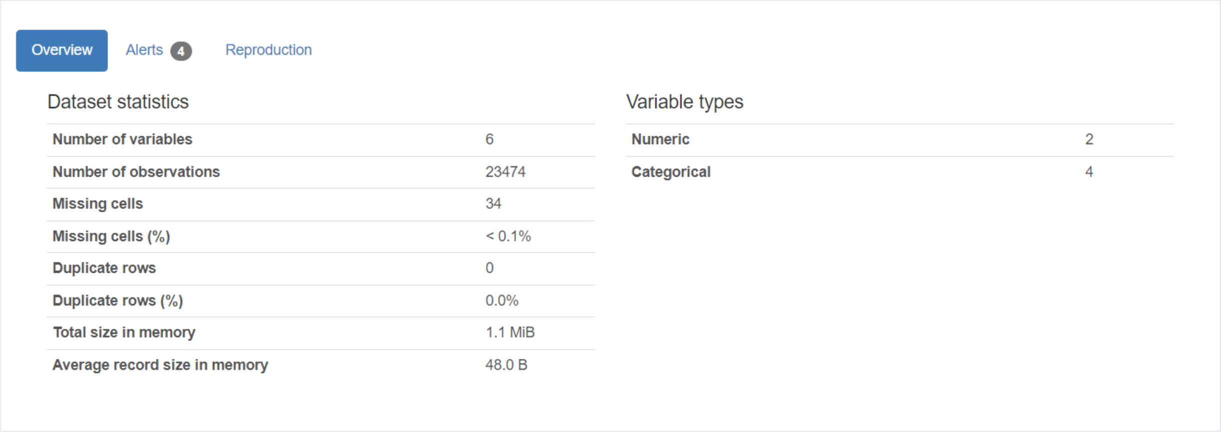
One-Hot Encoding:

1. Identifying the variables to be encoded.
2. Determine the number of categories in the variable.
3. Create a binary vector for each category, with a length equal to the number of categories.
4. For each observation in each dataset, find the category it belongs to.
5. One hot-encoded binary vectors are concatenated to create a new dataset with numerical features.

STEPS:

1. Analyzing and pre-processing data.

   Overview



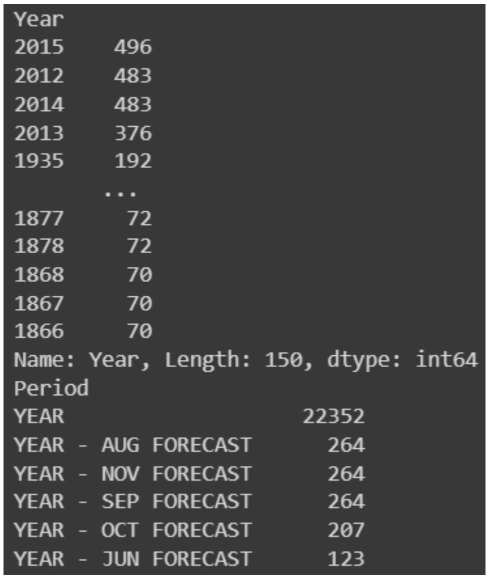
Dataset statistics

Number of variables	6
Number of observations	23474
Missing cells	34
Missing cells (%)	< 0.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.1 MiB
Average record size in memory	48.0 B

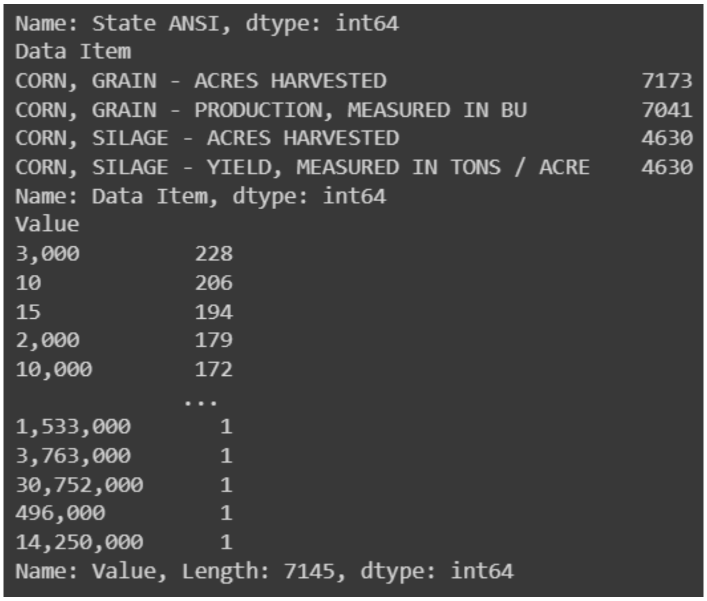
Variable types

Numeric	2
Categorical	4

2. Transforming Data to the required format. Here null values are removed to make a uniform dataset without discrepancies.



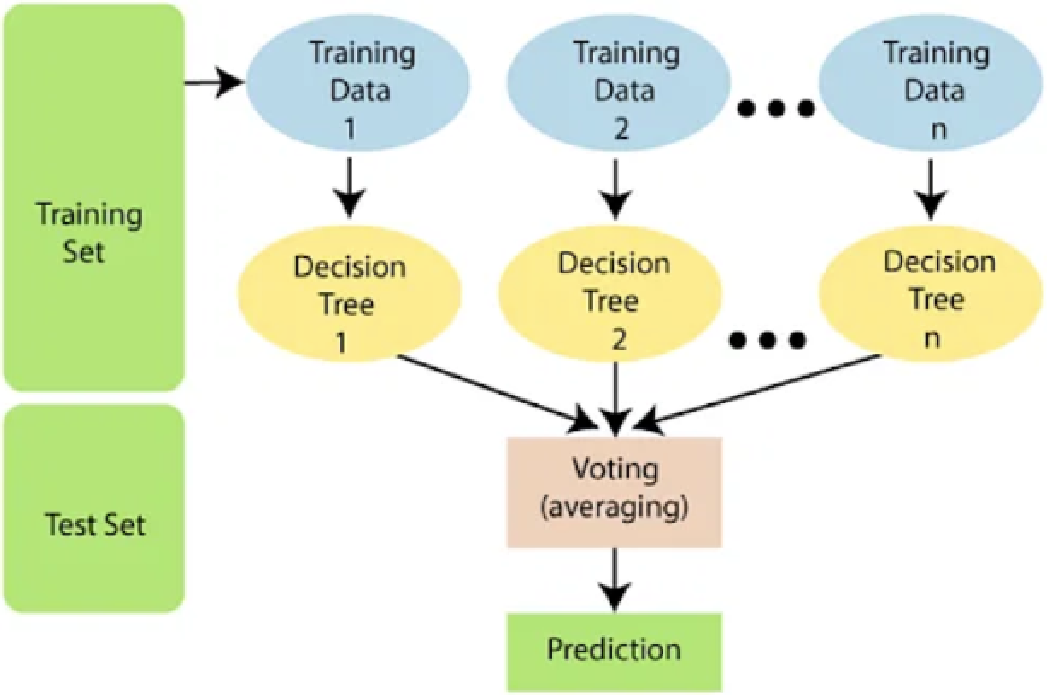
```
Year
2015    496
2012    483
2014    483
2013    376
1935    192
       ...
1877     72
1878     72
1868     70
1867     70
1866     70
Name: Year, Length: 150, dtype: int64
Period
YEAR                     22352
YEAR - AUG FORECAST        264
YEAR - NOV FORECAST        264
YEAR - SEP FORECAST        264
YEAR - OCT FORECAST        207
YEAR - JUN FORECAST        123
```



```
Name: State ANSI, dtype: int64
Data Item
CORN, GRAIN - ACRES HARVESTED                       7173
CORN, GRAIN - PRODUCTION, MEASURED IN BU            7041
CORN, SILAGE - ACRES HARVESTED                      4630
CORN, SILAGE - YIELD, MEASURED IN TONS / ACRE       4630
Name: Data Item, dtype: int64
Value
3,000         228
10            206
15            194
2,000         179
10,000        172
             ...
1,533,000       1
3,763,000       1
30,752,000      1
496,000         1
14,250,000      1
Name: Value, Length: 7145, dtype: int64
```

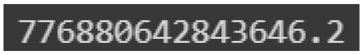
3. Applying One-hot encoding.

4. Comparing different models, which gives us results with regard to this dataset. The Random Forest Regressor comes out to be on top with an R^2 value of 0.9729. Knowing this is the most efficient method we will train it to find the predictions.



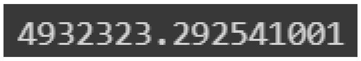
5. Training and Evaluating random forest metrics. From the findings we can see how close the values are thus proving the accuracy.

MSE Value:



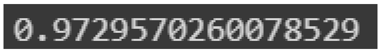
```
776880642843646.2
```

MAE Value:



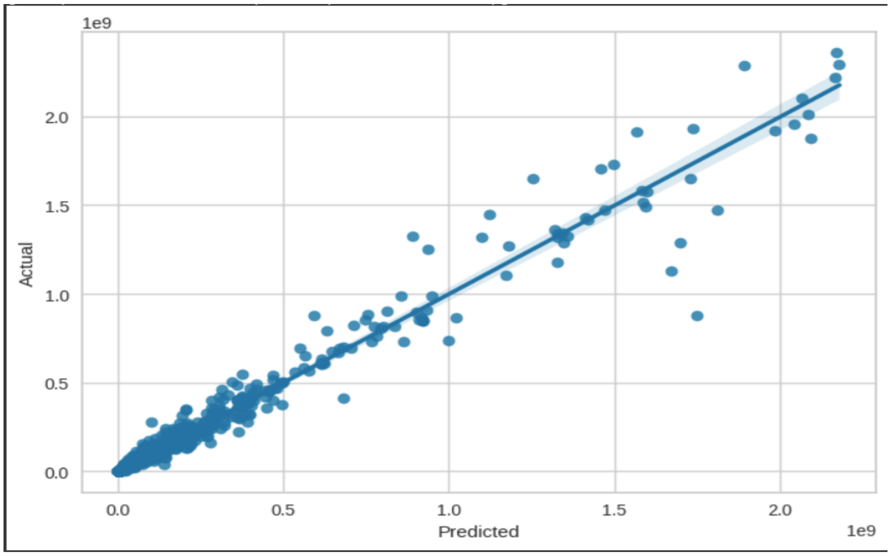
```
4932323.292541001
```

R^2 Value:

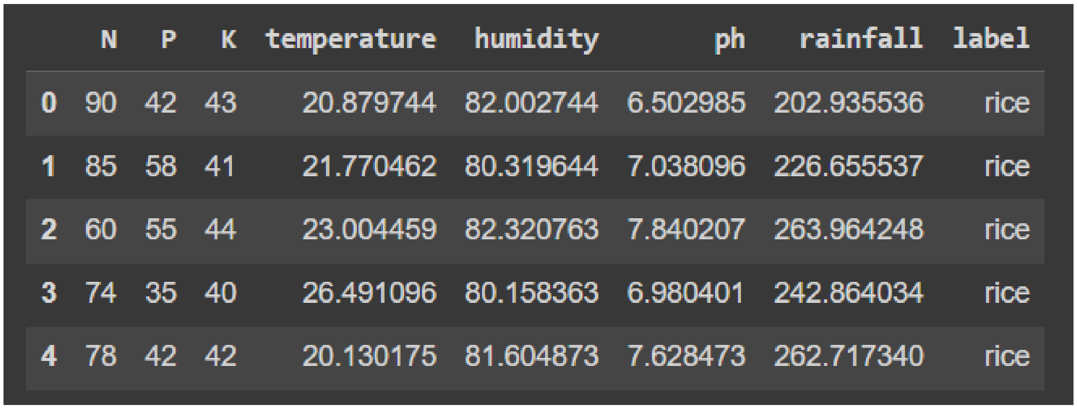


```
0.9729570260078529
```

Final Prediction vs. Actual Price Chart:

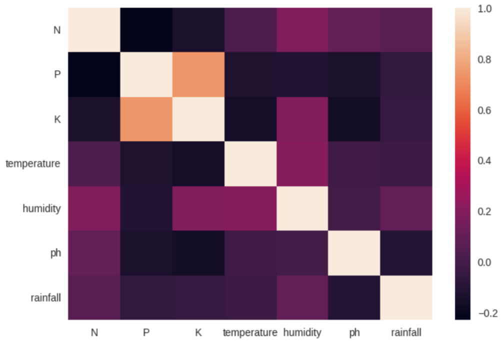


6. Analyze the second dataset.

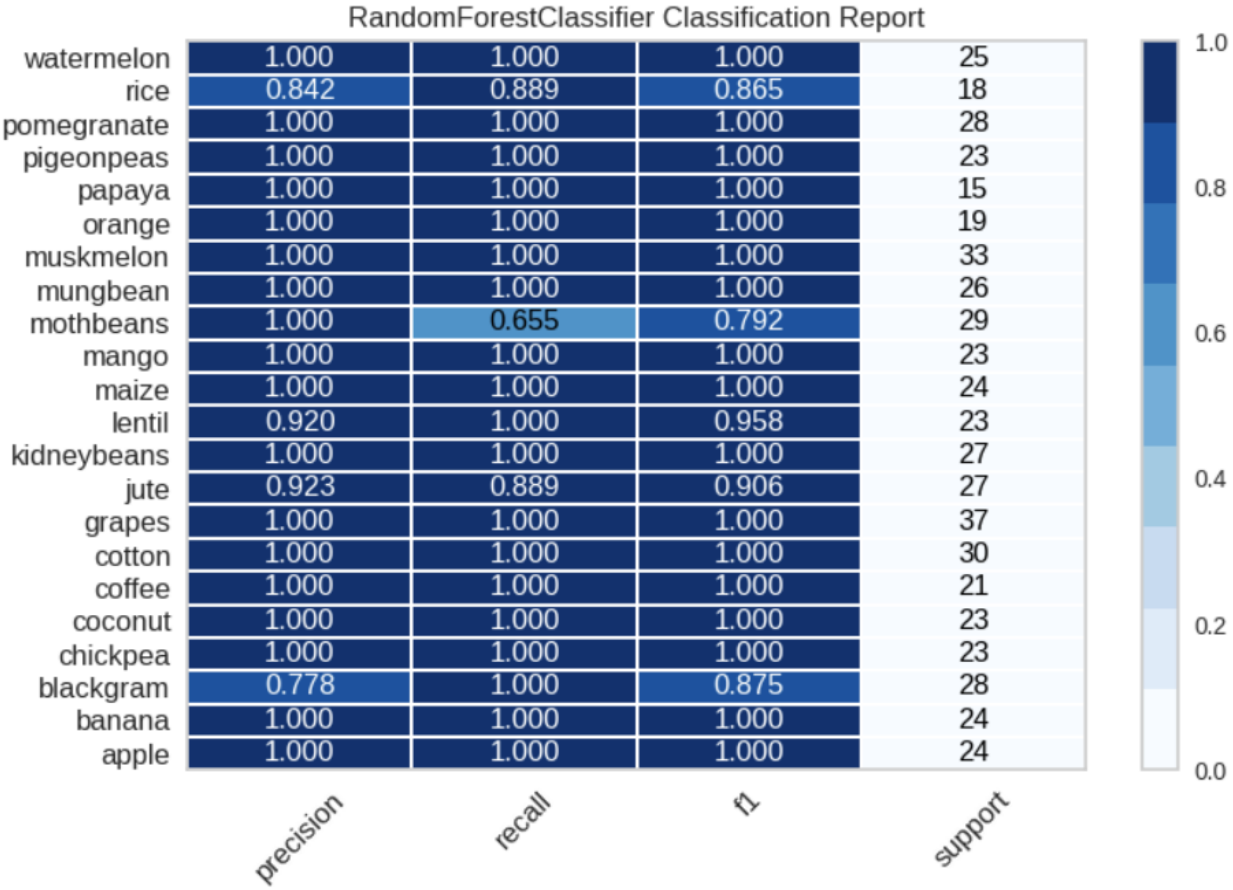


	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

7. Use heatmap to find correlation between provided parameters.



8. Used yellowbricks library to view the classification report of the random forest regressor on the dataset.



RandomForestClassifier Classification Report

	precision	recall	f1	support
watermelon	1.000	1.000	1.000	25
rice	0.842	0.889	0.865	18
pomegranate	1.000	1.000	1.000	28
pigeonpeas	1.000	1.000	1.000	23
papaya	1.000	1.000	1.000	15
orange	1.000	1.000	1.000	19
muskmelon	1.000	1.000	1.000	33
mungbean	1.000	1.000	1.000	26
mothbeans	1.000	0.655	0.792	29
mango	1.000	1.000	1.000	23
maize	1.000	1.000	1.000	24
lentil	0.920	1.000	0.958	23
kidneybeans	1.000	1.000	1.000	27
jute	0.923	0.889	0.906	27
grapes	1.000	1.000	1.000	37
cotton	1.000	1.000	1.000	30
coffee	1.000	1.000	1.000	21
coconut	1.000	1.000	1.000	23
chickpea	1.000	1.000	1.000	23
blackgram	0.778	1.000	0.875	28
banana	1.000	1.000	1.000	24
apple	1.000	1.000	1.000	24